Identification of gene clusters in rice genome

Project Report

Submitted in partial fulfilment of the requirements for the award of the

Degree of

MASTER OF SCIENCE
IN
BIOTECHNOLOGY

Submitted By

A.M. Nooman Siddique Barbhuiya

Reg.- 11309718

Under the guidance of

Dr. Atul Kumar Upadhyay

Assistant Professor

TO

DEPARTMENT OF BIOTECHNOLOGY AND BIOSCIENCES

LOVELY PROFESSIONAL UNIVERSITY

PHAGWARA, PUNJAB-144411

ACKNOWLEDGEMENT

I am greatly indebted to our beloved Pro Chancellor, Mrs. Rashmi Mittal and Vice Chancellor, Dr. Ramesh Kanwar for providing us with a healthy environment that drives us to achieve our ambitions and goals.

I am very thankful to my guide Dr. Atul Kumar Upadhyay for his invaluable guidance, assistance, patience and for giving this opportunity to explore the field of bioinformatics without which the accomplishment of the task would have not been probable.

I would also like to thank my lecturers and friends who were always there to give me a piece of advice and words of inspiration.

(**A.M. Nooman Siddique Barbhuiya**)

DECLARATION

I hereby declare that the project work entitled "**Identification of gene clusters in rice genome**" is an authentic record of my own work carried out at Lovely Professional University as requirement of project work for the award of Dessertation 1, under the guidance of Dr. Atul Kumar Upadhyay.

A.M. Nooman Siddique Barbhuiya

Reg.- 11309718

CERTIFICATE

This is to certify that **"A.M. Nooman Siddique Barbhuiya"** has done the project work entitled **"Identification of gene clusters in rice genome"** for the award of Dessertation 1 from "**Lovely Professional University", Phagwara, Punjab** under my supervision. This project embodies result of original study and research carried out by the student himself and the content of the project does not form the basis of any other degree to the candidate or anybody else.

Dr. Atul Kumar Upadhyay
Assistant Professor,
Department of
Biotechnology and Biosciences,
Lovely Professional University.

Table of contents

List of tables

List of figures

Introduction:

Rice (*Oryza sativa*) is of family Poaceae and genus *oryza* with over 20 cultivated wild species. *Oryza sativa* and *oryza glaberrima* being the most cultivated rice species. *Oryza sativa* is grown worldwide while *oryza glaberrima* has been cultivated for about last 3500 years in West Africa. Rice contains a basic chromosome number of n = 12. The species' can be diploid as well as triploid. Both *Oryza sativa* and *oryza glaberrima L.* being diploid species (2n = 24). Asian cultivated rice is the first fully sequenced crop genome. Rice is the staple food of more than 3 billion people worldwide. Significant quantity of recommended niacin and zinc are provided by rice. The digestibility of rice protein is very high (88 per.) and thus is considered as biologically richest protein. After wheat, rice is considered second in the most important crops of the world. (Ghosh et al.,2016).

Rice cultivation covers about 9 per. of the earth's cultivable land. Energy provided by rice to global human per capita is 21 per. and capital protein is 15 per.. In Asia, calories from rice are China, Indonesia and India the major producers as such Asia holds 90 per. of the world's rice production. Percentage (per.) of worlds rice crop traded in the world market is only 6-7 per.. (Boyer et. al., 2011).

World's largest exporters of rice are Thailand, China, Vietnam and United States. 1.5 per. of the world's rice crop is produced by United States with Arkansa, California and Louisiania producing 80 per. of the U.S rice crop production. Rice used for direct human consumption in the world is 85 per. (Boyer et. al., 2011). As rice is grown in different environmental conditions using different production methods, different rice varieties have different characteristics which make one variety of a particular area more popular than another variety. According the size of the grain, rice can be short, medium or long grain size. Rice can also be waxy (sticky) or non-waxy. The colour of the rice grain also varies, which can be brown, red, white, purple and black (Boyer et. al., 2011).

A gene family is an arrangement of homologous qualities inside one living being. A gene cluster is a bit of a gene family. A gene group is a get-together of no less than two qualities found inside a living being's DNA that encode for equivalent polypeptides, or proteins, which offers as a

summed up work and are as often as possibly arranged inside two or three thousand base arrangements of each other. The degree of gene cluster can move basically, from two or three qualities to a couple of hundred qualities.(Saito, 2009) Bits of the DNA progression of each quality inside a gene cluster are seen to be unclear; regardless, the resulting protein of each gene is specific from the consequent protein of another gene inside the pack.(Genes found in a gene cluster may be seen practically like each other on a comparable chromosome or on different. An instance of a gene cluster is the Hox gene, which contains eight genes and is a bit of the Homeobox quality family.(Field et al., 2011) Confirmations are available where it is conceivable to foresee useful coupling in view of protection of gene cluster between genomes. (Lawrence and Jeffrey, 1999).

Plants make a massive group of arranged metabolites that secure them against diseases, and abiotic stress. Plant discretionary metabolites choose harvest characteristics and are moreover a rich wellspring of bioactives for sedate and agrochemical disclosure. Around 20% of plant qualities have foreseen limits in discretionary processing, however even in Arabidopsis thaliana, the components of the dominant part of these are so far darken. The amount of sequenced plant genomes now outperforms 30. (Chen et.al., 2011; Sakakibara and Saito, 2009).

This collection of data is developing exponentially as sequencing techniques are created. Nonetheless, the metabolically diverse variety of plants remains generally unexploited, reason being the complexity of plant genome.

The finding that the qualities for the mix of different huge groups of plant-decided optional metabolite are made in gatherings, reminiscent of the operons and metabolic quality bunches found in microorganisms, is currently opening up new open doors for pathway disclosure (Chu et al.,, 2011; Osbourn, 2010; Osbourn and Field, 2009). In spite of the fact that these bunches have operon-like highlights (physical bunching and coregulation), they are unmistakably particular from bacterial operons in light of the fact that the genes inside each group are freely translated and isolated by huge extends of noncoding DNA. The main case of a plant metabolic gene cluster was for the combination of the cyclic hydroxamic acids 2,4-dihydroxy-1,4-benzoxazin-3-one (DIBOA) and 2,4-dihydroxy-7- methoxy-1,4-benzoxazin-3-one (DIMBOA) in maize (Frey

et al., 1997). It has been found and portrayed a further three groups, one in oat and two out of A. thaliana (for the combination of various triterpenes) (Field et al., 2011; Field and Osbourn, 2008; Qi et al., 2006). for different sorts of plant optional metabolite have too been accounted for from differing plant species. These gene clusters incorporate the phytocassane also, momilactone diterpenes in rice (Shimura et.al., 2007; Wilderman et.al., 2004). Including cyanogenic glucosides in Lotus japonicas andcassava (Takos et al., 2011).

Thus finding gene clusters will enable us to assign functions to uncharacterized genes in sequenced genomes. In this project we are identifying the gene clusters of Rice (*Oryza sativa*).

Materials and Method:

Softwares used in this project are **Plantismash** and **Blastp.**

Plantismash:

It allows the fast far reaching recognizable proof, explanation and examination of optional metabolite biosynthesis quality bunches over the plant kingdom. It is a particular augmentation of the generally utilized anti-SMASH instrument, customized particularly to target plant genomes.

In this software we first give the email to which we want our results to be sent. In the second input box we upload our file of genomic or nucleotide sequence in GenBank or EMBL format. The file as a whole is quite big in size. To avoid and cut off time consumption we can use the third input box. In this we upload the NCBI accession no. of our desired file. Another option would be to segment whole genome into their chromosomes and upload each individually. This helps in reducing time consumption significantly.

Once the results are obtained we can analyse them for finding gene clusters, their size, location and core domains.

Blastp:

As the name implies, BLAST(Basic Local Alignment Search Tool) does local alignments. Most proteins are secluded in nature, with at least one useful area happening inside a protein. Similar areas may likewise happen in proteins from various species. The BLAST calculation is tuned to discover these spaces or shorter extends of grouping closeness. The nearby arrangement approach likewise implies that a mRNA can be lined up with a bit of genomic DNA, as is much of the time required in genome gathering and examination. BLASTP performs protein-protein arrangement correlation, and its calculation is the premise of numerous different sorts of BLAST quests.

In this software we upload our protein sequence first in FASTA format. Or we can upload our sequence file directly. Then input the needed parameters (if) such as blast against a specific

organism database information, exclude an organism informations. We then need to select the algorithm for our search such as quickblast, phi blast among others. Click on blast.

The results will get displayed which we can further analyse for informations such as superfamilies, similar sequences, conserved regions, similarity tree among others.

Results:

Gene clusters of *Oryza sativa* (chr-1 to 12)

	Sr. No.	Gene cluster	Size (kb)	Core Domains
Chromosome 1	1.	Saccharide	71.44	2OG-FeII_Oxy, DIOX_N, Peptidase_S10, UDPGT_2
	2.	Lignan-Polyketide	70.90	Chal_sti_synt_C, Chal_sti_synt_N, Dirigent, p450
	3.	Saccharide	82.51	Aminotran_1_2, UDPGT_2
	4.	Saccharide	72.22	UDPGT_2, p450
	5.	Alkaloid	33.28	Bet_v_1, Epimerase, Methyltransf_11
Chromosome 2	1.	Saccharide	139.97	Glycos_transf_1, p450
	2.	Saccharide-Polyketide	211.17	Chal_sti_synt_C, UDPGT_2, p450
	3.	Terpene	369.98	COesterase, Terpene_synth, Terpene_synth_C, p450
Chromosome 3	1.	Lignan-Saccharide	97.55	Cellulose_synt, Dirigent, Methyltransf_11, UDPGT_2
	2.	Saccharide	64.12	Amino_oxidase, UDPGT_2, adh_short
Chromosome 4	1.	Terpene	212.71	Terpene_synth, Terpene_synth_C, adh_short_C2, p450
	2.	Saccharide-Alkaloid	360.51	Cu_amine_oxid, UDPGT_2, adh_short
	3.	Saccharide	169.20	Peptidase_S10, UDPGT_2

	4.	Terpene	334.35	2OG-FeII_Oxy, Terpene_synth, Terpene_synth_C
	5.	Saccharide	42.28	Peptidase_S10, UDPGT_2
	6.	Terpene	61.50	Terpene_synth, Terpene_synth_C, Transferase
	7.	Lignan	82.15	2OG-FeII_Oxy, DIOX_N, Dirigent, Methyltransf_7
Chromosome 5	1.	Saccharide	207.12	2OG-FeII_Oxy, DIOX_N, Transferase, UDPGT_2
Chromosome 6	1.	Putative	71.58	2OG-FeII_Oxy, DIOX_N
	2.	Putative	105.71	Peptidase_S10, Transferase, adh_short_C2
	3.	Saccharide	165.15	Transferase, UDPGT_2
	4.	Polyketide	133.31	Chal_sti_synt_C, p450
Chromosome 7	1.	Lignan	86.46	Aminotran_1_2, Dirigent
	2.	Lignan-Saccharide	88.18	Aminotran_1_2, Dirigent, Glycos_transf_1
	3.	Lignan	86.37	COesterase, Dirigent, p450
Chromosome 8	1.	Saccharide-Terpene	127.02	Methyltransf_2, Terpene_synth, Terpene_synth_C, UDPGT_2
	2.	Lignan-Alkaloid	132.28	Bet_v_1, Dirigent, Epimerase

	3.	Putative	83.82	COesterase, adh_short
Chromosome 9	1.	Saccharide	99.62	AMP-binding, UDPGT_2, p450
	2.	Putative	150.15	COesterase, Peptidase_S10, adh_short
Chromosome 10	1.	Saccharide	141.74	Transferase, UDPGT_2, p450
	2.	Lignan-Saccharide	432.20	Dirigent, UDPGT_2, p450
	3.	Polyketide	141.94	Acetyltransf_1, COesterase, Chal_sti_synt_C, Epimerase
	4.	Polyketide	139.12	Amino_oxidase, Chal_sti_synt_C, GMC_oxred_C, GMC_oxred_N
Chromosome 11	1.	Alkaloid	41.98	HMGL-like, Str_synth, p450
	2.	Lignan	130.12	Dirigent, Peptidase_S10
	3.	Saccharide	468.44	2OG-FeII_Oxy, UDPGT_2, adh_short, adh_short_C2
Chromosome 12	1.	Lignan	323.86	Dirigent, Methyltransf_2, p450
	2.	Saccharide	67.79	Glycos_transf_1, p450

Table 1

Results obtained have the most occurrence of saccharide and putative gene clusters in chromosomes (1 to 6 and 8 to 12) and (6,8,9) respectively.

Analysed results of putative:

2OG_fell_oxy:

Obtained Sequences of 2OG-Fe(II) oxygenase domain which contains protein, prediction via ab initio support : EST, InterPro domain: (IPR005123)

Sequence 1:

"MAMRSFVGDGGRVGGLRRRRQVVAVWDYGGGGGGGGEQPIPFRGNGGVLRQYLSYI
RMEGSMEDVRSTLLVQELAGMRSKSVPRQYIVQQEDQPTIAATASFPIVDLGRLSQPDG
DANEAVKLRQAMESWGLFMVTNHGIEDALMDNVMNVSREFFQQHLGEKQKYTNLIDG
KHFQLEGYGNDQVKSDTQILDWLDRLYLKVDPADERNLSVWPKHPESFRDVLDEFLIK
CDGVKNSLLPSMAKLLKLNEDYFVRQFSDRPTTIARFNYYPQCPRPDLVYGMKPHSDAT
ILTILMVDNDVGGLQVLKDGVWYDVPTKPHTLLINLGDHMELLLKLTVSL"

Sequence 2:

"MAMRSFVGDGGRVGGLRRRRQVVAVWDYGGGGGGGGEQPIPFRGNGGVLRQYLSYI
RMEGSMEDVRSTLLVQELAGMRSKSVPRQYIVQQEDQPTIAATASFPIVDLGRLSQPDG
DANEAVKLRQAMESWGLFMVTNHGIEDALMDNVMNVSREFFQQHLGEKQKYTNLIDG
KHFQLEGYGNDQVKSDTQILDWLDRLYLKVDPADERNLSVWPKHPESFRDVLDEFLIK
CDGVKNSLLPSMAKLLKLNEDYFVRQFSDRPTTIARFNYYPQCPRPDLVYGMKPHSDAT
ILTILMVDNDVGGLQVLKDGVWYDVPTKPHTLLINLGDHMELLLKLTVSL"

Sequence 3:

"MADGHHWNIVKIPPIVQELAAGVHEPPSQYMVGEKDRPAIAGSDMPEPIPVVDLSRLSA
SNGEDSAGELAKLRSALEDWGLFLGSILSEMINVTRGFYKLPLEEKQKYSNLVNGKDFRI
EGYGNDMVVSEKQILNWEITSLVLARLARLLGLREGYFVDMFDEDATTYARFNYYPRC
LRPEDVLGLKPHSDGSVITVVSVDDTVSGLQVLRQGVWYDVPVVPNALLINMGDGMEI
MSNGLLKSPVHRVVTNAERERVSVVMFYALDPEKELEPAPELVDDEKRPRQYAKMKIK

DYLSGFYETFARGTRVIDTVKMSE"

Sequence 4:

"MVHPAQGQMVQDLAAGGELGAPPSRYVLREKDRPVAAAGAVQAAQRELAAIPTIDVS
RLAAESGDDVVDDGGEAAKLRSALQSWGLFAVTGHGMPEPFLDEILAATREFFHLPPEE
KERYSNVVAAADADGVGAGGERFQPEGYGIDRVDTDEQILDWCDRLYLQVQPEEERRL
EFWPEHPAALRGLLEEYTRRSEQVFRRVLAATARSLGFGEEFFGDKVGEKVTTYARFTY
YPPCPRPELVYGLKPHTDNSVLTVLLLDKHVGGLQLLKDGRWLDIPVLTNELLVVAGDE
IELFALLGVADHEQVFMAPVHRVVTSERERMSVVMFYQPEPHKELAPSEELVGEERPA
MY"

Sequence 5:

"MAMRSFVGDGGRVGGLRRRRQVVAVWDYGGGGGGGGEQPIPFRGNGGVLRQYLSYI
RMEGSMEDVRSTLLVQELAGMRSKSVPRQYIVQQEDQPTIAATASFPIVDLGRLSQPDG
DANEAVKLRQAMESWGLFMVTNHGIEDALMDNVMNVSREFFQQHLGEKQKYTNLIDG
KHFQLEGYGNDQVKSDTQILDWLDRLYLKVDPADERNLSVWPKHPESFRDVLDEFLIK
CDGVKNSLLPSMAKLLKLNEDYFVRQFSDRPTTIARFNYYPQCPRPDLVYGMKPHSDAT
ILTILMVDNDVGGLQVLKDGVWYDVPTKPHTLLINLGDHMELLLKLTVSL"

Sequence 6:

"MADGHHWNIVKIPPIVQELAAGVHEPPSQYMVGEKDRPAIAGSDMPEPIPVVDLSRLSA
SNGEDSAGELAKLRSALEDWGLFLGSILSEMINVTRGFYKLPLEEKQKYSNLVNGKDFRI
EGYGNDMVVSEKQILNWEITSLVLARLARLLGLREGYFVDMFDEDATTYARFNYYPRC
LRPEDVLGLKPHSDGSVITVVSVDDTVSGLQVLRQGVWYDVPVVPNALLINMGDGMEI
MSNGLLKSPVHRVVTNAERERVSVVMFYALDPEKELEPAPELVDDEKRPRQYAKMKIK
DYLSGFYETFARGTRVIDTVKMSE"

Out of all the sequences sequence 1 was used for obtaining results.

Similarity tree obtained in putative cluster one of chromosome 6:

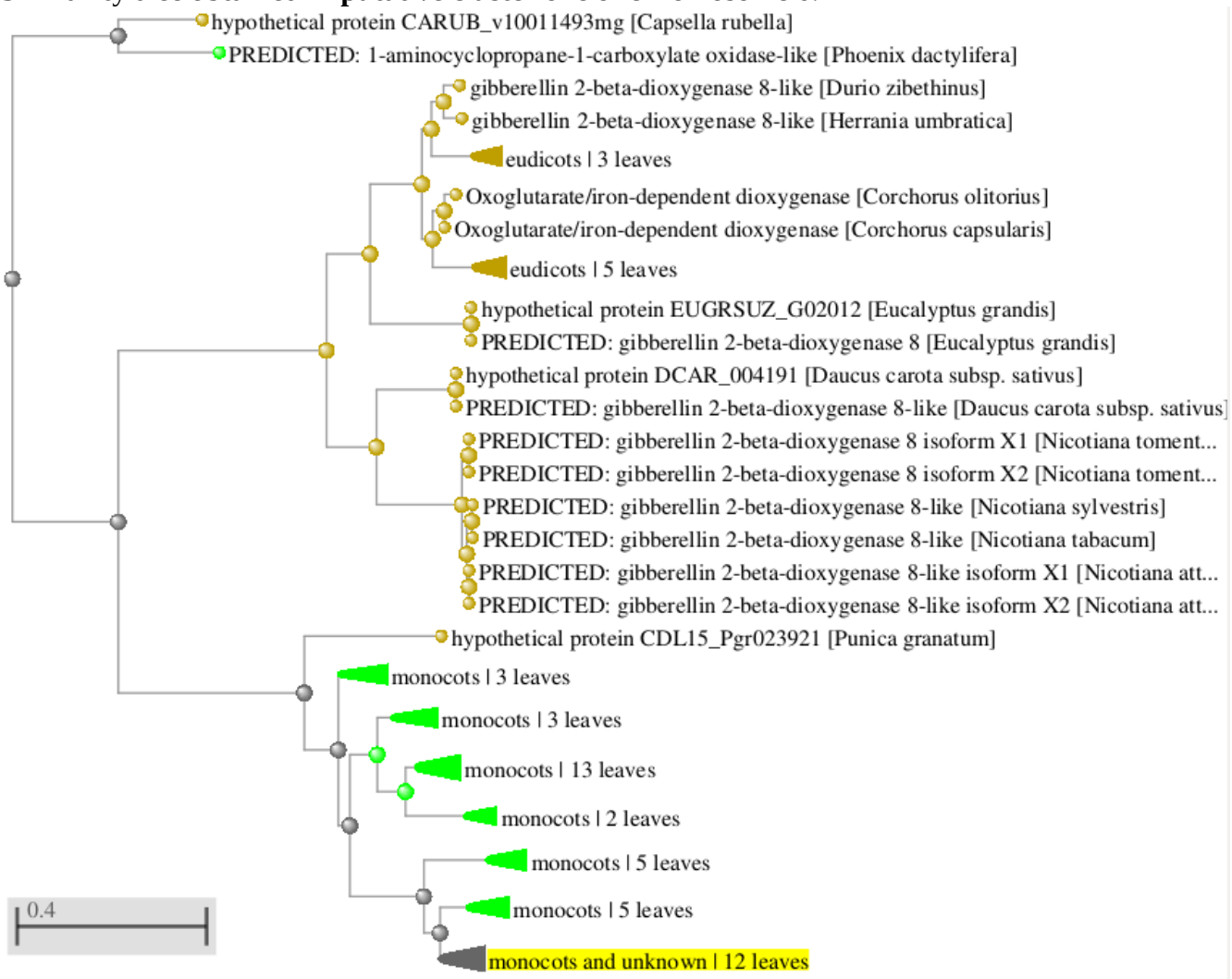

Fig.1

Out of 26 similar sequences that were obtained in the tree, predicted sequences obtained were 9. The most similarity was exohibited by the hypothetical protein CARUB. Least were exhibited by the predicted proteins.

Superfamilies of putative cluster one of chromosome 6:

Name	Accession	Description	Interval	e-value
PLN00417	PLN00417	oxidoreductase, 2OG-Fe(II) oxygenase family protein	51-352	5.15e-89
PcbC	COG3491	Isopenicillin N synthase and related dioxygenases	50-347	3.29e-33
DIOX-N	pfam14226	non-haem dioxygenase in morphine synthesis N-terminal	51-180	3.00e-26

Table 2

Name	Accession	Description	Interval	e-value
PLN00417	PLN00417	oxidoreductase, 2OG-Fe(II) oxygenase family protein	16-311	1.37e-74
2OG-FeII-Oxy	pfam03171	2OG-Fe(II) oxygenase superfamily	170-263	1.86e-28
PcbC	COG3491	Isopenicillin N synthase and related dioxygenases	44-306	6.53e-26

Table 3

Conclusion:

The gene clusters obtained via plantismash of rice chromosomes 1-12 were varied in each with quite a few similar clusters present in each. The predominant cluster that was obtained was saccharide. Putative has second most occurrence.

Core domains which have most occurrence are 2OG-FeII_Oxy, transferases and peptidases.

2OG-FeII_Oxy : Proteins with Fe and 2-oxoglutarate (2OG) dioxygenase space typically catalyze the oxidation of a characteristic substrate using a dioxygen molecule, generally by using ferrous iron as the dynamic site cofactor and 2OG as a co-substrate which is decarboxylated to succinate and CO2. In rice, FeII-2OG dioxygenase space impetuses catalyze the advancement of plant hormones like gibberellin, ethylene, hues and flavones.

Transferases : Methyltransferases and UDPGT transferases were the most prevalent transferases that were observed from obtained results. Uridine 5'- diphospho-glucuronosyltransferase (UDP-glucuronosyltransferase, UGT) exists as a cytosolic glycosyltransferase which does the transfer of the glucuronic acid part of Uridine 5'- diphospho-glucuronic acid to a little hydrophobic atom. This is a glucuronidation response. Methyltransferases as the name suggests transfers Methyl group from one functional group to another.

Peptidases : Peptidase_S10 most observed peptidase in obtained results, this cluster of serine peptidases(SPep) belong to MEROPS peptidase family S10 (group SC). All known carboxypeptidases are serine carboxypeptidases(SCPep.) or metallo carboxypeptidases. The reactant action of the SCPep. is similar to that of the trypsin family SP, is given by a transfer framework including aspartic acid buildup hydrogen-attached to a histidine, which is itself hydrogen-linked to a serine.

References:

Boyer, J. S. (1982). Plant productivity and environment. *Science,* 218(4571), 443-448.

Chen, S., Xiang, L., Guo, X., & Li, Q. (2011). An introduction to the medicinal plant genome project. Frontiers of Medicine, 5, 178–184.

Chu, H.-Y., Wegel, E., & Osbourn, A. (2011). From hormones to secondary metabolism: The emergence of secondary metabolic gene clusters in plants. The Plant Journal, 66, 66.

Field, B., & Osbourn, A. E. (2008). Metabolic diversification—Independent assembly of operon-like gene clusters in different plants. Science, 320, 543–547.

Field, B., Fiston-Lavier, A. S., Kemen, A., Geisler, K., Quesneville, H., & Osbourn, A. E. (2011). Formation of plant metabolic gene clusters within dynamic chromosomal regions. Proceedings of the National Academy of Sciences of the United States of America, 108, 16116–16121.

Frey, M., Chomet, P., Glawischnig, E., Stettner, C., Grun, S., Winklmair, A., et al. (1997). Analysis of a chemical plant defense mechanism in grasses. Science, 277, 696–699.

Ghosh, B., Ali, M.N. & Gantait, S. (2016). Response of Rice under Salinity Stress: A Review Update. *J. Res. Rice,* 4(2): 2–9.

Lawrence, Jeffrey (1999). Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Current Opinion in Genetics & Development*. 9(6): 642–8.

Osbourn, A., & Field, B. (2009). Operons. Cellular and Molecular Life Sciences, 66, 37555–37575.

Osbourn, A. (2010). Secondary metabolic gene clusters: Evolutionary toolkits for chemical innovation. Trends in Genetics, 26, 449–457.

Qi, X., Bakht, S., Qin, B., Leggett, M., Hemmings, A., Mellon, F., et al. (2006). A different function for a member of an ancient and highly conserved cytochrome P450 family: From essential sterol to plant defense. Proceedings of the National Academy of Sciences of the United States of America, 103, 18848–18853.

Shimura, K., Okada, A., Okada, K., Jikumaru, Y., Ko, K. W., Toyomasu, T., et al. (2007). Identification of a biosynthetic gene cluster in rice for momilactones. The Journal of Biological Chemistry, 282, 34013–34018.

Takos, A. M., Knudsen, C., Lai, D., Kannangara, R., Mikkelsen, L., Motawia, M. S., et al. (2011). Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in Lotus japonicus and suggests the repeated evolution of this chemical defence pathway. The Plant Journal, 68, 273–286.

Wilderman, P. R., Xu, M., Jin, Y., Coates, R. M., & Peters, R. J. (2004). Identification of Syn - imara-7,15-diene synthase reveals functional clustering of terpene synthases involved in rice phytoalexin/allelochemical biosynthesis. Plant Physiology, 135, 2098–2105.

Yonekura-Sakakibara, K., & Saito, K. (2009). Functional genomics for plant natural product synthesis. Natural Product Reports, 26, 1466–1487.